Received 13 June 2022, accepted 20 July 2022, date of publication 8 August 2022, date of current version 11 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3194545

RESEARCH ARTICLE

Missing Data Imputation Algorithm for Transmission Systems Based on Multivariate Imputation With Principal Component Analysis

YEON-SUB SIM[1], **JAE-SANG HWANG**[2], **SUNG-DUK MUN**[2], **TAE-JOON KIM**[2], **AND SEUNG JIN CHANG**[1]

[1]Department of Electric Engineering, Hanbat National University, Daejeon 34158, South Korea
[2]Korea Electric Power Corporation Research Institute, Daejeon 34056, South Korea

Corresponding author: Seung Jin Chang (jpromo8@gmail.com)

This work was supported in part by the Korea Electric Power Corporation under Grant R20TA12, and in part by the National Research Foundation under Grant NRF-2021R1A2C1095779.

ABSTRACT As the importance of utility condition is increasingly acknowledged, the use of asset management technologies in the electric power industry has rapidly grown. The global trend of asset management follows risk management that accounts for the probability and consequences of failures. Because asset management systems tend to be composed of various legacy systems, each of which is installed and designed to collect data according to a certain data type and acquisition purpose, it is necessary to develop a system that cleans and integrates data acquired from each legacy system. This study explores the development of an asset management system for a transmission system as a representative linear asset consisting of different segments in a sequence. First, the configurations and characteristics of linear asset datasets are analyzed. Second, an automatic data cleaning system, equipped with six types of data cleaning functions for extracting dirty data from entire datasets, is proposed. An algorithm for data imputation, which is essential for estimating the remaining useful life, is developed based on principal component analysis–iterative algorithm (PCA–IA). Afterward, the performance of the proposed system is verified using actual data with the help of the Korea Electric Power Corporation (KEPCO). Specifically, to evaluate the performance of the proposed system, an automatic cleaning process is demonstrated using actual legacy datasets.

INDEX TERMS Transmission system, data cleaning, database management, data imputation, principal component analysis, linear asset, machine learning.

I. INTRODUCTION

Thus far, the convergence of information and communication technology (ICT) and power systems has driven the maintenance of power systems to evolve from time-based maintenance (TBM) to condition-based maintenance (CBM). Alternatively, in recent years, prognosis-based maintenance (PBM) of power systems with the help of machine learning (ML) techniques has been explored in research studies. In any case, because of the diversification of energy sources, including of renewable energy, and the complexity of power grids, the reliability of power systems has become an even more important concern. One task of the power system operator is to determine the optimal replacement timing by considering the remaining useful life, damages from failure, and replacement costs of assets. The risk-based replacement priority method, which is implemented using a risk matrix consisting of the probability of failure (PoF) and consequence of failure (CoF), has been used to reduce maintenance costs. For this reason, many power system operators across the world have already started using asset management systems (AMS) to increase their business value [1].

The risk assessment algorithm is composed of six steps as follows: (1) data understanding, (2) data capture and analysis, (3) data quality evaluation, (4) data cleaning processing, (5) data quality re-evaluation, (6) data continual management.

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Raza.

The transmission system in power system plays a role in transmitting power over long distances, so a failure may cause a wide-area blackout with huge economic loss. For reliable transmission system operation, three types of legacy systems are installed and operated in South Korea according to handled data type as follows: (1) asset specification legacy system, (2) asset inspection and diagnosis legacy system, (3) loading information system [2]. A transmission system is a representative linear asset in which lines are connected to each other in order, and lines are composed of cables and joint boxes. A linear assets are defined by length with their length directly impacting their maintenance such as highways, pipelines. Because each basic element is connected, the failure of each part may affect the connected parts as well. It is necessary to propose a new basic asset unit to consider the ripple effect of the faulty part on the connected parts. In addition, various data collected in each legacy system must be integrated according to the asset connection order and new basic asset unit. Existing legacy systems have difficulties in data integration, because even identifiers for each system are not unified, and the collected data is often contaminated.

However, because incorrect legacy data can affect the results of replacement priority evaluations, good quality legacy data are necessary for accurate investment planning. Therefore, it is generally agreed that data cleaning processes are essential for asset management. It has also been reported that data scientists spend 60 % of their time on cleaning and organizing data [3]–[6]. Thus, data cleaning and integration tools are required to increase the reliability of AMS. The process of data cleaning can be divided into missing data and outlier detection, and data refining and imputation. Oftentimes, the data acquired by sensors will have missing points, and although it is not difficult to determine that parts of the data are missing, it is difficult to accurately estimate the values of the missing data. Data imputation refers to the process of estimating missing data based on observed data, and is actively being studied for its application in the medical field. Missing data can be classified into three types, as follows: (1) the missing data are completely random and independent of other variables, referred to as missing completely at random (MCAR); (2) the missing data are dependent on observation data, referred to as missing at random (MAR); and (3) the missing data are dependent on both observed and non-observed variables, referred to as missing not at random (MNAR). In this study, we develop a data imputation method and verify its performance under the assumption of MCAR [7].

The most traditional method of tackling missing data is to simply delete the dataset that contains the missing data. With regard to obtaining good-quality datasets, deleting a dataset that contains dirty data is better than inaccurately estimating the missing data; however, this method has the serious disadvantage of reducing the overall data size. This decrease in data size leads to insufficiencies in the training data used to derive information through a learning process. An alternative is to replace missing data with derived plausible values through a process known as missing data imputation. A variety of techniques for data imputation have been developed using statistical and machine learning approaches. State-of-the-art imputation methods include principal component analysis (PCA) [8], [9] based on machine learning, expectation maximization based on statistical methods, and autoencoders and generative adversarial nets (GAN) based on deep learning [10]–[14]. Each method is more applicable than the others in certain situations based on its advantages and disadvantages. For example, expectation maximization requires assumptions about data distribution and cannot be applied to a dataset with a mixture of continuous and categorical variables [15], whereas an autoencoder can be used for estimating missing data when part of the dataset is missing, but requires a complete dataset for training [16]. On the other hand, generative adversarial imputation nets (GAIN), the latest technique for data imputation based on GAN, exhibit excellent data imputation performance even when complete data are unavailable [10]–[14].

The transmission line data used in this study involve the following considerations. (1) Because failures of transmission lines have a huge economic ripple effect, few cases of failure are available for analysis. Moreover, only a few lines are installed in similar environments. In other words, it is difficult to secure the massive data required by existing deep learning methods for each group of cases because cases of replacements after a failure are presently insufficient. (2) The amount of real-time data acquired from transmission lines installed across the country of South Korea is quite large, and thus it is necessary to develop a lighter algorithm to enable control within 1 s in case of failure based on the data. On the other hand, various deep learning techniques, including GAIN and GAN, are used to manage a variety of devices, such as motors and robots, installed separately in a factory. The biggest differences between the factory setup and the transmission lines examined in this study is that the lines operate in a connected way, and thus the complexity of the algorithm can be lowered using domain knowledge relevant to the power industry. The ultimate purpose of the data management of transmission lines is to predict the remaining lifespan of a transmission line in operation based on the operation data of a faulty line, and to select the optimal asset replacement timing based on economic evaluation of assets in need of replacement. Based on the results of this study as a cornerstone, a remaining-life prediction and replacement-timing selection algorithm will be developed based on GAN, long short-term memory (LSTM), and optimization methods.

In this study, the asset data are categorized as numeric, categoric, and string data based on analyses of actual power asset data from across South Korea. Subsequently, algorithms for detecting outliers and dirty data according to contamination type are introduced. In the case of numeric data, an algorithm for replacing missing data is developed using machine learning. A system that repeats a series of processes until the data quality reaches a set value is then proposed.

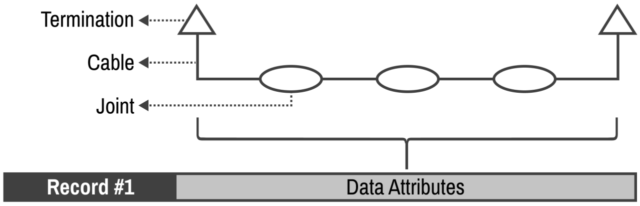


FIGURE 1. Data management when unit of linear asset is a circuit.

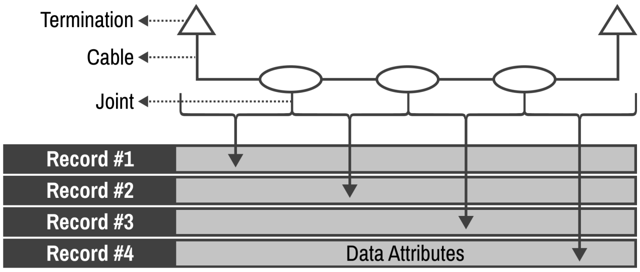


FIGURE 2. Data management when unit of linear asset is a segment.

In Section 2, we introduce an automatic cleaning system that includes the data characteristics of linear assets, three legacy systems, and six data cleaning functions. There are three types of legacy systems related to transmission systems, and thus the data acquired from each segment constituting the transmission system should be classified and the data integrated according to the segment connection order. The transmission system installed in South Korea consists of underground cables and overhead lines, of which the cable types are oil-filled (OF) and cross-linked polyethylene (XLPE) cables, which are representative of the cable types in South Korea. The automatic data cleaning system is equipped with six data cleaning functions, which are introduced further in the paper, and among them, a calculation function that processes data acquired from load information legacy systems is described in detail. We then examine the algorithm that handles missing data in the load information legacy system.

In Section 3, a case involving missing data found in a target linear asset is discussed. The utilization rate data of the load information legacy system, which we mainly estimated, are of a numeric data type, and in an analogy between power assets and humans, these data are similar to human workout data. We then propose an imputation algorithm that selects other linear assets that are similar to the target asset considering the asset connection order, and the missing data are replaced based on PCA–IA using the similar asset data.

Finally, in Section 4, our newly developed automatic legacy data cleaning and imputation system, constructed based on these algorithms, is described and verified using actual transmission cable data from South Korea.

II. DATA CLEANING ALGORITHM FOR TRANSMISSION SYSTEM

Linear assets are characterized as having a linear structure, where they are arranged in a row, with the components connected to each other serially. Because linear assets are interconnected, a failure in one part also affects the connected parts. To reflect these properties of a linear asset, we set the basic linear asset unit as one cable section and a joint box on both sides. This study explores the following power assets: (1) cable type: underground cable, overhead line, (2) rated voltage: 154 kV, 345 kV, (3) cable insulation type: XLPE, OF, and (4) asset type: cable, joint box, termination box.

Legacy Data and Systems of Transmission System: Linear assets are installed under different environmental conditions, given the interconnectedness of these assets among each other [17], [18]. In addition, the asset information characteristics, such as age, cable type, and installation environment, may be different among segments.

As shown in Fig. 1, setting the entire circuit as the basic asset unit is disadvantageous in terms of capital expenditure, because the entire circuit needs to be replaced even if only a part of the unit fails. By contrast, this paper proposes a method of inputting one record as a segment, as shown in Fig. 2. The joint box or termination box on both sides and the connecting cable are regarded as one basic asset unit. This basic unit composed of three assets makes it easier to monitor the health statuses of the connected parts in the event that one asset fails, and can prevent economic waste that would otherwise require replacing the entire circuit.

Legacy systems of transmission cable systems can be classified into legacy systems for asset specification, inspection and diagnosis, and loading information. These legacy systems are connected to each other using key ID data for interfacing. Among these legacy systems, information systems for assets are similarly classified into three types based on the type of handled and collected information: (1) asset specification legacy systems, (2) asset inspection and diagnosis legacy systems, and (3) loading information legacy systems.

An information system for asset specification manages the history of the overall transmission equipment, from data creation to destruction, based on geographic information. Representative data include cable type, circuit length, manufacturer, and date of installation. Asset specification data, which are akin to the year of birth, gender, etc., do not change and are, in another analogy to human identification data, similar to information on a registration card. Through the use of this information, it is possible to identify the unique characteristics of an asset that do not change over time.

On the other hand, information systems for asset inspection and diagnosis results are used to record data from annual inspections or special diagnoses. These systems are interfaced with information systems for asset specification. Representative data include diagnosis results on partial discharge, dissolved gas analysis (DGA) of insulating oil, and thermal hot spots. An inspection and diagnostic information system involves three types of cable diagnostic data and three types of joint box diagnostic data, depending on the subject. Through the diagnostic information, the health status of the power equipment can be monitored.

With regard to information systems for loading information, a variety of parameters, such as voltage, current, active power, reactive power, and utilization rate of cables, are recorded and managed. Through this loading information, it is possible to infer the remaining useful life of the subject.

A. DATA CLEANING ALGORITHM

Because the data cleaning work encompasses more than 60 % of the total process, it needs to be automated to improve data quality. Because a power asset has a long lifespan of more than 40 years, it is necessary to manage assets over a long term. Because the legacy systems of a transmission system were installed after most of the power assets were already in operation, all of the data from before the legacy system installation are missing. Hence, there is a need to develop a data cleaning system equipped with a data imputation algorithm. Data cleaning work includes the collection of data, detection of missing data, and classification of outlier data. Based on an analysis of the types of dirty patterns, we introduce six types of cleaning setting functions, which befit different cases: 1) transform, 2) pattern, 3) scanning, 4) historical, 5) criteria, and 6) calculation functions [6]. In the following sections, these cleaning functions are briefly introduced to explain the algorithm, and among them, the calculation function, which is important for estimating the states of assets, is described in further detail.

1) TRANSFORM FUNCTIONS

The transform function is used to convert data after a checking rule-based cleaning method is applied. It can be used for unified circuit names or manufacturer names. Although the circuit names should be unified according to the specified internal guidelines, mistyped data could occur because of human error. In the case of cable manufacturers, these data are often hand-typed. As a result, the same name may have different names according to individual style. In such cases, varying names such as "LS Cable," "LG Cable," "LS Cable System," and "LG Cable System" are cleaned to a unique name, "LS Cable."

2) PATTERN FUNCTIONS

The pattern function detects outlier data based on the data pattern. For example, an AC transmission system follows the form of a three-phase system consisting of A, B, and C phases. Although the numbers of A, B and C phases should be the same, there are times when the numbers of phases may not be identical because of human error. For example, "A, B, and B phase" or "A, B, and missing" could be automatically cleaned to "A, B, and C phase" using the pattern function after the rest of the information is checked for consistency with the information on the other phases.

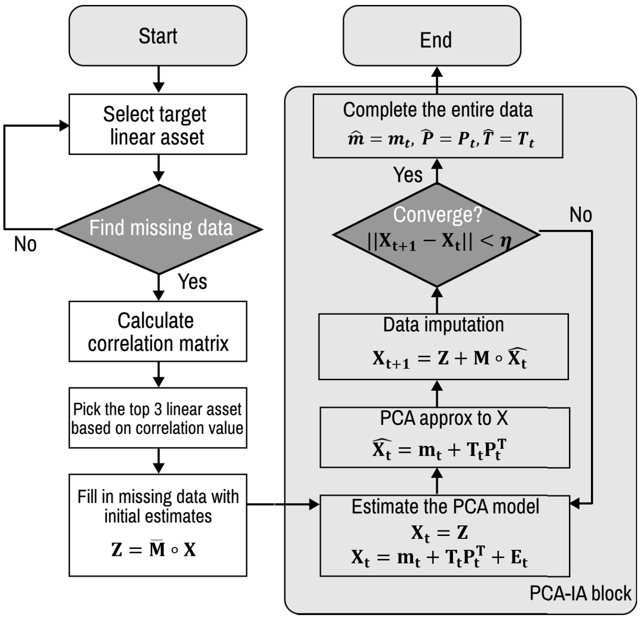

Start
End
Select target linear asset
Find missing data
No
Yes
Calculate correlation matrix
Pick the top 3 linear asset based on correlation value
Fill in missing data with initial estimates
$\mathbf{Z} = \overline{\mathbf{M}} \circ \mathbf{X}$
Complete the entire data
$\hat{m} = m_t,\ \hat{P} = P_t, \hat{T} = T_t$
Yes
Converge?
$||\mathbf{X}_{t+1} - \mathbf{X}_t|| < \eta$
No
Data imputation
$\mathbf{X}_{t+1} = \mathbf{Z} + \mathbf{M} \circ \widehat{\mathbf{X}_t}$
PCA approx to X
$\widehat{\mathbf{X}_t} = \mathbf{m}_t + \mathbf{T}_t\mathbf{P}_t^T$
Estimate the PCA model
$\mathbf{X}_t = \mathbf{Z}$
$\mathbf{X}_t = \mathbf{m}_t + \mathbf{T}_t\mathbf{P}_t^T + \mathbf{E}_t$
PCA-IA block

FIGURE 3. Flowchart of utilization rate data imputation.

3) SCANNING FUNCTIONS

The scanning function detects outlier data by checking the uniqueness of the data. Although each equipment must have a unique keycode, which is automatically created and assigned to distinguish them, duplicate keycodes for the same asset can be generated because of data redundancy. The best way to detect this outlier problem is to check the number of keycodes assigned to one circuit name. Through the confirmation of legacy system operators, the duplicate keycodes can then be revised into a unique keycode.

4) HISTORICAL FUNCTIONS

The historical function cleans missing or dirty data, which are to be replaced based on historical information for the base date. For example, since April 1, 2005, cable termination insulators have been changed from porcelain to polymer material. To properly reflect history, historical data can be automatically converted to refer to porcelain insulators if the points are dated up to March 31, 2005, and to polymer insulators if the points are dated from April 1, 2005 and beyond.

5) CRITERIA FUNCTIONS

The criteria function processes numeric data by analyzing its scatter data plot. Specifically, the data are displayed as a scatter plot, from which abnormal data can be detected based on the data distribution. In the following example, thermal inspection of cables and joint boxes is used as a representative case. The maximum temperature was measured on site using a thermal imaging camera, and the measurement data were uploaded. The measured temperature was 22 °C; however, it was incorrectly input to the legacy system as 222 °C. In this case, based on the entire dataset, the criteria function extracts the appropriate temperature range. Appropriate boundaries can be set automatically based on data distribution analysis, or can be set manually based on user opinions. Hence, the outlier data are extracted and cleaned.

6) CALCULATION FUNCTIONS

The calculation function calculates the utilization rate using the active/reactive power of the circuit. The utilization rate information, which is stored in a load information legacy system, is important for estimating the remaining lifetime, which will be explored in future studies. However, one of the problems involved herein is the occurrence of missing data on the utilization rate. In this case, the utilization rate data can be calculated using active/reactive power information, as follows (1):

$$U_{rate} = \frac{\sqrt{P^2[W] + Q^2[VAR]}}{\sqrt{3}V[V] * I[A]}, \qquad (1)$$

where U_{rate}, which is a utilization rate, can be derived using the calculation function based on active power, reactive power, and rated voltage, which are stored in the load information legacy system; and ampacity, which is stored in the information system for asset specification. That is, when the utilization rate data are missing, they can be calculated based on a combination of information stored in other legacy systems.

III. AUTOMATIC DATA IMPUTATION ALGORITHM

The most important and difficult problem in the data cleaning process is the data imputation of missing data. In the case of loading information that is measured once every hour by a load information legacy system, missing data cannot be calculated using the calculation function when the input data of the calculation function, such as active power and reactive power, are omitted. Because the loading data, including the utilization rate, active power, and reactive power, will be necessary for deriving the remaining useful life of power assets in the future, it is important to now develop an imputation method. Data handled by the loading information platform include utilization rate and reactive/active power, and the performance of the algorithm is verified by estimating active power data. To handle the missing data, we propose an imputation method based on the principal component analysis–iterative algorithm (PCA–IA) [19], [20].

A. DATA IMPUTATION RESULTS

Fig. 3 shows the entire proposed algorithm for data imputation of missing values. First, we select the target linear asset, and check whether there are missing data. If there are missing data, we calculate the correlation matrix between the target linear asset and the linear asset sharing the substation. Based on the characteristics of the linear asset connecting the starting and end points, then when there are missing data, the opposite system data must be checked first. The characteristics of linear assets that are connected to each other result in utilization rate data that are similar between the data of assets sharing the substation. The correlation matrix is derived by the following equation, which is an indicator of how well data are related to each other.

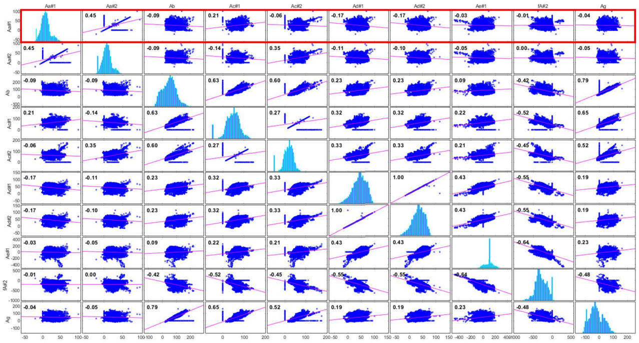


FIGURE 4. Correlation matrix for sharing of common substation.

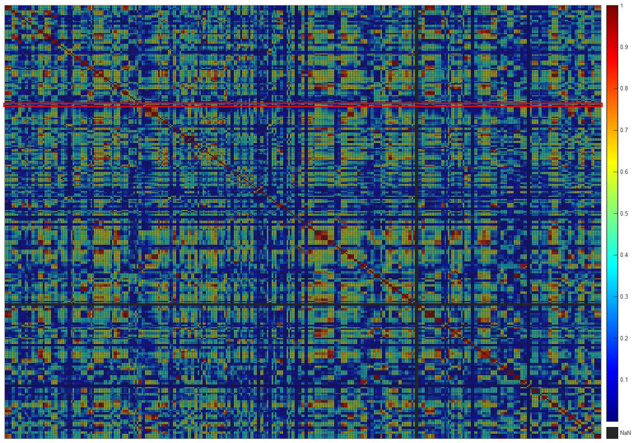


FIGURE 5. Entire correlation matrix.

$$S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1M} \\ s_{21} & s_2^2 & \cdots & s_{2M} \\ \vdots & & \ddots & \\ s_{M1} & s_{M2} & \cdots & s_M^2 \end{pmatrix} \qquad (2)$$

where, $s_{jk} = (1/M)\sum_{i=1}^{M}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})$ is the co-variance between the j-th and k-th variables, $s_j^2 = (1/M)\sum_{i=1}^{M}(x_{ij} - \overline{x_j})^2$, is the variance of the j-th variable, and $\overline{x_j} = (1/M)\sum_{i=1}^{M} x_{ij}$ is the mean of the j-th variable. In general, when the value exceeds 0.5, it can be determined that there is a significant relationship. If, through correlation analysis, it is determined that there are no similar data among the assets sharing the substation, it will be necessary to determine whether there is an asset with high correlation among the total asset data. Based on this, we select the top three linear assets that have the highest correlation value. The missing data are first filled with the average value of the data of the target asset as initial value, and then restored through PCA–IA. With the entire dataset and the initial estimate value, $Z = \overline{M} \circ X$. Matrix M is an indicator matrix of missing data, whereas matrix Z contains the original data without missing values. We estimate the PCA model as $X_t = m_t + T_t P_t^T + E_t$, where t is the current time, and m_t and E_t are the average and the measurement error of observations, respectively.

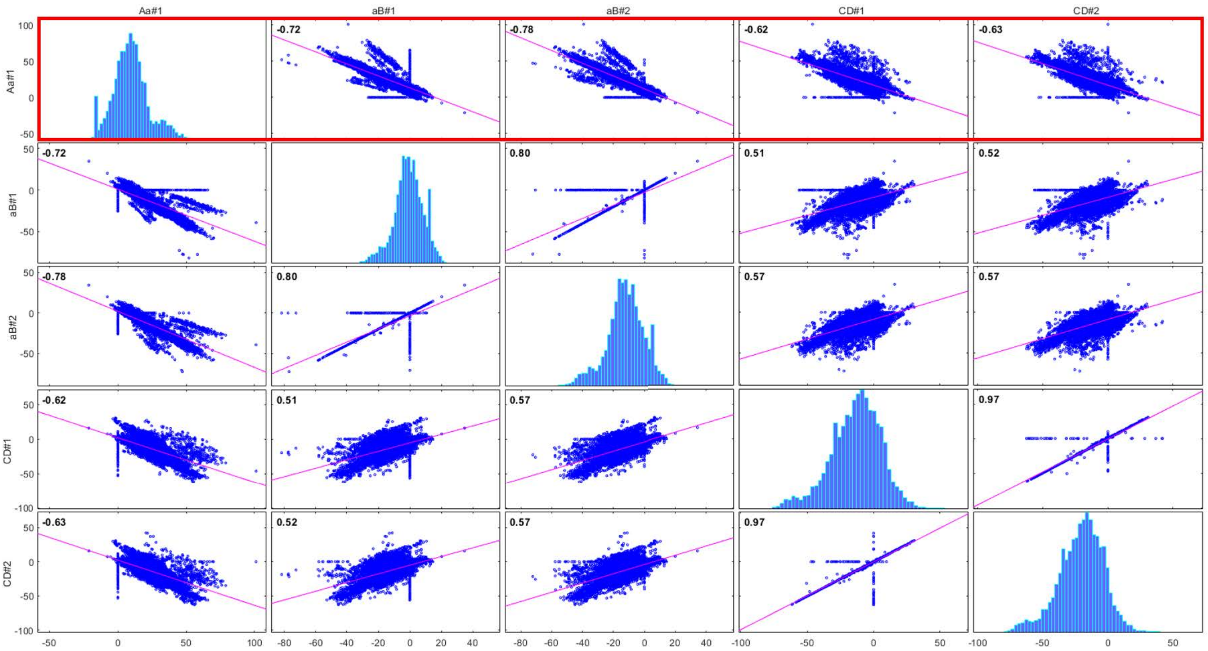


FIGURE 6. Target linear assets with highest correlation coefficient value.

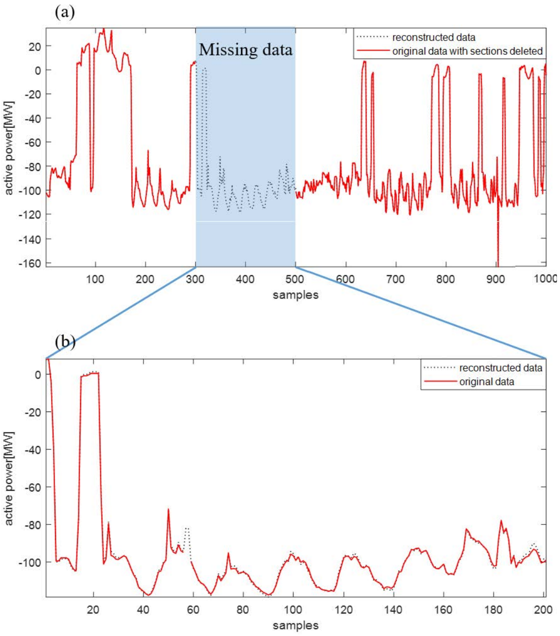


FIGURE 7. Scenario I. (a) comparative analysis of data imputation, and (b) results of data imputation algorithm.

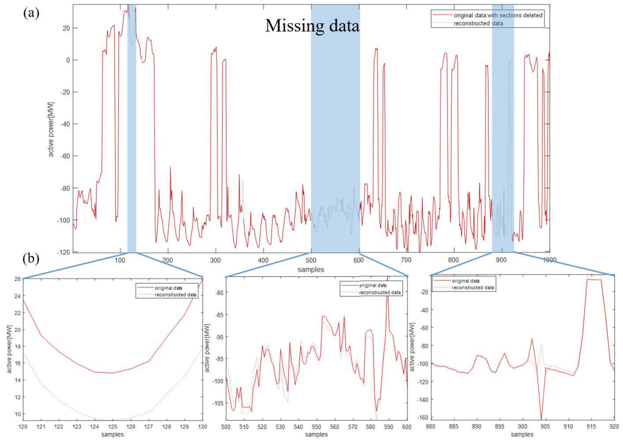
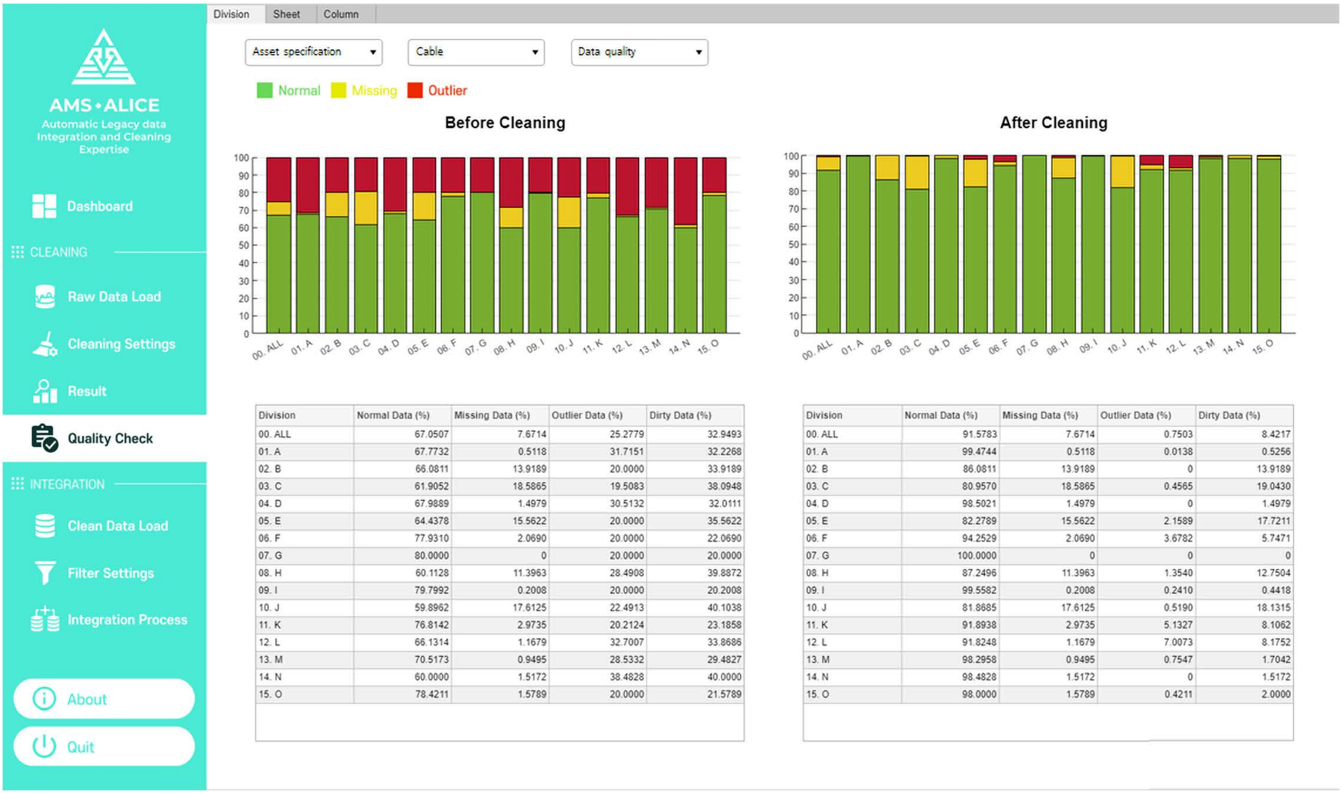


FIGURE 8. Scenario II. (a) comparative analysis of data imputation, and (b) results of data imputation algorithm.

We consider the data matrix, X, $M \times K$ matrix, which denotes K different time zones and M different objects. For example, $x_{i,j}$ represents the j_{th} data point acquired from the i_{th} object. This signifies that the i_{th} row of the matrix contains data from the i_{th} object, and is denoted by x_i^T and the full matrix. PCA is based on decomposition of the data matrix X into two orthogonal matrices, V and U. Data matrix X is represented by $X = TP^T$, where T is an $M \times K$ matrix of scores, and P is a $K \times K$ matrix of loadings. The loadings are the weights for the variables. Each observation in the new coordinate system of principal components is called a score, which is calculated as a linear combination of the original variables and loadings [18]. Without loss of generality, the missing data can be positioned at the first element of the data vector, such that the vector can be partitioned as follows: $X = \left[X^* X^{\circledast}\right]$, where X^* and $X^{\circledast}$ are the missing and measured data, respectively; X^* is the submatrix containing the first R columns of data matrix X, and $X^{\circledast}$ contains the remaining $N-R$ columns. Likewise, the loading matrix P can be partitioned as $P = \left[P^* P^{\circledast}\right]$, where P^* is the submatrix containing the first R rows of loading matrix P, and $P^{\circledast}$ contains the remaining $N - R$ rows. Based on the constructed PCA model, the entire dataset, including



Division	Normal Data (%)	Missing Data (%)	Outlier Data (%)	Dirty Data (%)
00. ALL	67.0507	7.6714	25.2779	32.9493
01. A	67.7732	0.5118	31.7151	32.2268
02. B	66.0811	13.9189	20.0000	33.9189
03. C	61.9052	18.5865	19.5083	38.0948
04. D	67.9889	1.4979	30.5132	32.0111
05. E	64.4378	15.5622	20.0000	35.5622
06. F	77.9310	2.0690	20.0000	22.0690
07. G	80.0000	0	20.0000	20.0000
08. H	60.1128	11.3963	28.4908	39.8872
09. I	79.7992	0.2008	20.0000	20.2008
10. J	59.8962	17.6125	22.4913	40.1038
11. K	76.8142	2.9735	20.2124	23.1858
12. L	66.1314	1.1679	32.7007	33.8686
13. M	70.5173	0.9495	28.5332	29.4827
14. N	60.0000	1.5172	38.4828	40.0000
15. O	78.4211	1.5789	20.0000	21.5789

Division	Normal Data (%)	Missing Data (%)	Outlier Data (%)	Dirty Data (%)
00. ALL	91.5783	7.6714	0.7503	8.4217
01. A	99.4744	0.5118	0.0138	0.5256
02. B	86.0811	13.9189	0	13.9189
03. C	80.9570	18.5865	0.4565	19.0430
04. D	98.5021	1.4979	0	1.4979
05. E	82.2789	15.5622	2.1589	17.7211
06. F	94.2529	2.0690	3.6782	5.7471
07. G	100.0000	0	0	0
08. H	87.2496	11.3963	1.3540	12.7504
09. I	99.5582	0.2008	0.2410	0.4418
10. J	81.8685	17.6125	0.5190	18.1315
11. K	91.8938	2.9735	5.1327	8.1062
12. L	91.8248	1.1679	7.0073	8.1752
13. M	98.2958	0.9495	0.7547	1.7042
14. N	98.4828	1.5172	0	1.5172
15. O	98.0000	1.5789	0.4211	2.0000

FIGURE 9. Comparison of data quality check before and after cleaning.

missing data, is estimated. Only the parts corresponding to missing data from the entire estimated dataset are added with the initially estimated data to update the data as $X_{t+1} = Z + M \circ \hat{X}_t$, where Z is equal to X_0, which is an initial estimate value. Through replacement of the original missing data with the values predicted from the PCA, the model is fitted. In this study, PCA modeling is performed based on the three linear asset data with the highest relationship, as obtained through correlation analysis. Because PCA transform is a kind of linear transformation, the model error converges to the bounded output and does not diverge. The PCA process is iterated until convergence of the predicted values for the missing data. The threshold, η, can be set by the user with consideration for the required time and desired accuracy.

IV. DEVELOPMENT OF DATA CLEANING AND INTEGRATION SYSTEM

An automatic data cleaning system for asset management was developed based on both cleaning and imputation algorithms, as described in Section III. After the development of this system, it is demonstrated using loading information legacy datasets utilized for the asset management of electric power equipment in South Korea.

To handle the missing data for the loading information legacy system, we propose an algorithm based on PCA–IA. Because linear assets are connected to each other, they exhibit high correlation values between assets sharing the same substation. Fig. 4 shows the correlation matrix for linear assets sharing a common substation. Marked on the x-axis and y-axis are the name of the linear asset, where the first uppercase letter refers to the substation, and the second lowercase letter refers to the line name. It can be observed that the linear assets share the same substation. The diagonal is the waveform to be compared, and the remaining waveforms are the correlation results. According to the characteristics of linear assets, most loading data have high correlation values with other assets that share the same substation; however, the red boxed data show very low correlation values (0.4 or less) with other connected assets. In this case, the correlation value between the target asset and the entire asset with which it does not share a substation should be analyzed, as shown in Fig. 5. The bar displayed on the right side of Fig. 4 is a color representation of the correlation value. Based on the entire correlation matrix, we can select the top three linear assets that have the highest correlation coefficient, as shown in Fig. 6. The missing data of the target asset can be estimated by using the three selected lines in Fig. 6. To verify the performance of the proposed method, we established two scenarios according to the length of the dropout time: (1) a signal in which 20 % of the total signal length is omitted, and (2) a signal in which less than 5 % of the total signal length is missing multiple times (120 − 130, 500 − 600, 880 − 920 samples). We compared the reconstructed signal with the original signal after deleting the data. As shown in Fig. 7(a), we intentionally deleted 20 % of the active power data extracted from the loading information legacy. Fig. 7(b) enlarges the missing section of Fig. 7(a), and compares the signal reconstructed based on the proposed method with the original signal before omission. Based on the comparative analysis of the reconstructed signal against the original signal before omission, it can be confirmed that the signal imputation performance is excellent. Fig. 8(a) shows the estimated data for when data for a short period of time are simultaneously missing, whereas Fig. 8(b) shows an enlarged figure. It can be observed that a time delay occurs in estimating the part where the active power changes rapidly, although this estimation can be achieved accurately within a certain amount of time using data of other selected lines.

A. DEMONSTRATION EXPERIENCE

Through the proposed automatic cleaning process, the cleaning time can be drastically reduced to within approximately one day, compared to the manual method, which requires several months. According to the results of the data cleaning demonstration, the legacy data accuracy increased from approximately 70 % to over 91 %, as shown in Fig. 9 Green, yellow, and red bars on the UI/UX shown in Fig. 9 indicate the normal, missing, and contaminated data, respectively. In addition, the results of evaluating the data quality for 15 divisions in South Korea are shown in the bar graphs, and it can be verified from Fig.9 that the performance improved after the data cleaning process. Because cleaning a set of data with average or regression values without the verification of a manager may lead to different results, it is difficult to clean the data using only an automatic cleaning algorithm. For this reason, the legacy system manager has to manually check the cleaned data values derived by the automatic cleaning algorithm.

The data imputation method presented in this paper is applicable to the data cleaning part of a platform for data cleaning and integration. We have developed a platform equipped with the proposed method, and it is now used for the data management of transmission lines installed in all regions of South Korea. Moreover, the proposed algorithm is not limited to being used on transmission systems and can be used practically anywhere. The target asset of this study is a linear asset that is connected to other such assets. Because the proposed system utilizes the characteristics of connectivity, it works differently from managing the data of devices, such as motors and robots, inside a factory. Rather, (1) there is a connection between assets, and (2) where numerical data obtained from various sensors can be utilized, the proposed method can be used immediately. For example, the proposed algorithm can be applied to data management for condition monitoring using sensors attached to pipelines and highways.

V. CONCLUSION

Herein, a novel data management system for managing the data of transmission systems is proposed. This data management system is divided into three parts: 1) data cleaning, 2) data imputation, and 3) evaluation of data quality. This paper mainly introduces an algorithm based on PCA–IA that replaces missing data in load information legacy systems. Its basic asset unit is composed of the cable section and the joint box at both ends. The cleaning part consists of six functions distinguished according to data characteristic, and the set values are modified through the incorporation of expert opinions. The imputation algorithm is designed for the loading information, given that it is the most important information for the prediction of remaining useful life, which is to be studied further in future research endeavors. The cleaned data are sent to each legacy system, which then collects data for feedback. The performance of the automatic cleaning algorithm gradually improves through this feedback. A system for evaluating the data quality of each system at each regional office was also constructed, to evaluate the data quality before and after cleaning using actual power equipment data from all over South Korea, and to verify the performance of the proposed system through the feedback of the managers of each system. The data management system proposed in this paper is expected to become a touchstone for the development of upcoming remaining-lifetime evaluation systems for power assets.

REFERENCES

[1] M. Mc Granaghan, "Making connections: Asset management and the smart grid," *IEEE Power Energy Mag.*, vol. 8, no. 6, pp. 16–22, 2015.

[2] J.-S. Hwang, S.-D. Mun, T.-J. Kim, G.-W. Oh, Y.-S. Sim, and S. J. Chang, "Development of data cleaning and integration algorithm for asset management of power system," *Energies*, vol. 15, no. 5, p. 1616, Feb. 2022.

[3] I. F. Ilyas and X. Chu, *Data Cleaning*. New York, NY, USA: Association for Computing Machinery, 2019.

[4] *Data Science Report*, CrowdFlower, San Francisco, CA, USA, 2016.

[5] M. Balazinska, A. Deshpande, M. J. Franklin, P. B. Gibbons, J. Gray, M. Hansen, M. Liebhold, S. Nath, A. Szalay, and V. Tao, "Data management in the worldwide sensor web," *IEEE Pervasive Comput.*, vol. 6, no. 2, pp. 30–40, Apr. 2007.

[6] X. Chu, "Data cleaning: Overview and emerging challenges," in *Proc. Int. Conf. Manage. Data*, 2016, pp. 2201–2206.

[7] S. Nakagawa and R. P. Freckleton, "Missing inaction: The dangers of ignoring missing data," *Trends Ecol. Evol.*, vol. 23, no. 11, pp. 592–596, 2008.

[8] J. X. Mi, Q. Zhu, and J. Lu, "Principal component analysis based on block-norm minimization," *Appl. Intell.*, vol. 49, no. 6, pp. 2169–2177, 2019.

[9] M.-J. Wu, "Integrative hypergraph regularization principal component analysis for sample clustering and co-expression genes network analysis on multi-omics data," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1823–1834, 2019.

[10] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," *Proc. Mach. Learn. Res.*, vol. 80, pp. 5689–5698, 2018.

[11] J. Kim, D. Tae, and J. Seok, "A survey of missing data imputation using generative adversarial networks," in *Proc. Int. Conf. Artif. Intell. in Inf. Commun. (ICAIIC)*, 2020, pp. 454–456.

[12] W. Zhang, P. Zhang, Y. Yu, X. Li, S. A. Biancardo, and J. Zhang, "Missing data repairs for traffic flow with self-attention generative adversarial imputation net," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7919–7930, May 2021.

[13] X. Liu and Z. Zhang, "A two-stage deep autoencoder-based missing data imputation method for wind farm SCADA data," *IEEE Sensors J.*, vol. 21, no. 9, pp. 10933–10945, May 2021.

[14] Z. Pan, Y. Wang, K. Wang, H. Chen, C. Yang, and W. Gui, "Imputation of missing values in time series using an adaptive-learned median-filled deep autoencoder," *IEEE Trans. Cybern.*, early access, May 4, 2022, doi: 10.1109/TCYB.2022.3167995.

[15] P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, 2017.

[16] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[17] Y. Sun, C. Fidge, and L. Ma, "Reliability prediction of long-lived linear assets with incomplete failure data," in *Proc. Int. Conf. Quality, Rel., Risk, Maintenance, Saf. Eng.*, Jun. 2011, pp. 143–147.

[18] Y. Sun, L. Ma, W. Robinson, M. Purser, A. Mathew, and C. Fidge, "Renewal decision support for linear assets," in *Proc. 5th World Congr. Eng. Asset Manag. (WCEAM AGIC)*, 2010, pp. 885–899.

[19] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg, "Active-Clean: Interactive data cleaning while learning convex loss models," 2016, *arXiv:1601.03797*.

[20] P. R. C. Nelson, P. A. Taylor, and J. F. Macgregor, "Missing data methods in PCA and PLS: Score calculations with incomplete observations," *Chemometrics Intell. Lab. Syst.*, vol. 35, no. 1, pp. 45–65, 1996.

YEON-SUB SIM was born in Daejeon, South Korea. He received the B.S. degree from the Department of Electrical Engineering, Hanbat National University, Daejeon, in 2016.

His research interests include asset management systems based on machine/deep learning, efficient data analysis for complex data, and related applications.

JAE-SANG HWANG received the B.S. and the combined M.S./Ph.D. degrees from the Department of Electronic Systems Engineering, Hanyang University, South Korea, in 2011 and 2016, respectively.

Since 2015, he has been with the Korea Electric Power Corporation (KEPCO) Research Institute, Daejeon, South Korea, where he is currently a Senior Researcher with the Asset Management System Research Team. His research interests include high-voltage insulation, electric field analysis, data preprocessing, risk assessment algorithms for transmission cable systems, and asset management systems.

SUNG-DUK MUN received the B.S. and M.S. degrees from the Department of Electronic Systems Engineering, Hanyang University, South Korea, in 2013 and 2015, respectively.

Since 2016, he has been with the Korea Electric Power Corporation (KEPCO) Research Institute, Daejeon, South Korea, where he is currently a Researcher with the Asset Management System Research Team. His current research interests include asset management systems, end-of-lifetime, and electromagnetics for overhead transmission lines.

TAE-JOON KIM received the B.S. and M.S. degrees from the Department of Polymer Engineering, Chonnam University, Gwangju, South Korea, in 2011 and 2013, respectively.

In 2014, he joined the Korea Institute of Industrial Technology, Bucheon, South Korea, as a Researcher. Since 2019, he has been a Researcher with the Korea Electric Power Research Institute, Daejeon, South Korea. His current research interests include risk- based asset management systems for transmission cables and lines.

SEUNG JIN CHANG was born in Seoul, South Korea. He received the B.S. and Ph.D. degrees from the Department of Electrical and Electronic Engineering, Yonsei University, Seoul, in 2010 and 2017, respectively.

In 2018, he joined the School of Electrical and Computer Engineering, Seoul National University, as a Postdoctoral Researcher. Since 2018, he has been an Assistant Professor with the Department of Electrical Engineering, Hanbat National University, Daejeon, South Korea. His current research interests include condition monitoring based on machine/deep learning, diagnosis and prognostics of power equipment including cables and batteries, applied signal processing techniques, and time–frequency analysis.